Conditional Mutual Information Based Boosting for Facial Expression Recognition

Caifeng Shan, Shaogang Gong and Peter W. McOwan
Department of Computer Science
Queen Mary, University of London, London E1 4NS, UK
{cfshan, sgg, pmco}@dcs.qmul.ac.uk

Abstract

This paper proposes a novel approach for facial expression recognition by boosting Local Binary Patterns (LBP) based classifiers. Low-cost LBP features are introduced to effectively describle local features of face images. A novel learning procedure, Conditional Mutual Information based Boosting (CMIB), is proposed. CMIB learns a sequence of weak classifiers that maximize their mutual information about a candidate class, conditional to the response of any weak classifier already selected; a strong classifier is constructed by combining the learned weak classifiers using the Naive-Bayes. Extensive experiments on the Cohn-Kanade database illustrated that LBP features are effective for expression analysis, and CMIB enables much faster training than AdaBoost, and yields a classifier of improved classification performance.

1 Introduction

Automatic facial expression recognition has attracted much attention [5, 14] in recent years. Though much progress has been made [4, 3, 2, 13], recognizing facial expression with a high accuracy remains to be difficult due to the complexity and variety of facial expressions. Facial expression recognition involves two vital aspects: facial feature representation and classifier design. Facial feature representation is to derive a set of features from original face images which minimizes within-class variations of expressions whilst maxmizes between-class variations. If inadequate features are used, even the best classifier could fail to achieve accurate recognition. There are two common approaches to extract facial features: geometric feature-based methods and appearance-based methods [14]. Gabor-wavelet appearance features were demonstrated to be more effective than geometric features [17], and more robust in low-resolution facial expression recognition [13]. In Donato et al's experiments [4], Gabor wavelet representation also performed best. Although Gabor-wavelet representations have been widely adopted [17, 2, 13], it is computationally expensive to convolve face images with multi-banks of Gabor filters in order to extract multiscale and orientational coefficients.

Local Binary Patterns (LBP) were proposed originally for texture analysis [11]. Recently Ahonen et al [1, 8] presented LBP based face detection and recognition, where the facial area is equally divided into small regions to extract LBP features. However, the

LBP features extracted from equally divided sub-regions suffers from fixed size and positions. By shifting and scaling a sub-window over face images, much more features could be obtained, which yield a more complete description of face images. For the very large number of LBP features introduced by shifting and scaling a sub-window, boosting learning [7] can be utilised to learn the most effective LBP features and boost weak classifiers to a strong classifier.

In this work, we first exploit Local Binary Patterns as low-cost discriminative appearance features for facial expression recognition (Section 2). Our motivation is that face images can be seen as a composition of micro-patterns which can be effectively described by LBP. Compared to Gabor wavelets, LBP features can be derived very fast in a single scan of raw images, whilst still retaining enough facial information in a compact representation. We then utilize boosting learning to learn a small set of optimal LBP features from a very large LBP feature pool. In addition to AdaBoost (Section 3), we further propose a novel learning procedure, Conditional Mutual Information based Boosting (CMIB), to boost LBP-based weak classifiers for improved expression recognition (Section 4). CMIB enables efficient learning of a sequence of weak classifiers by maximising their mutual information about a candidate class, conditional to the response of any weak classifier already selected, thus avoiding the selection of ineffective weak classifiers. In Section 5, extensive experiments using the Cohn-Kanade database show that LBP features are effective for expression analysis, and CMIB outperforms AdaBoost in boosting LBP features for expression recognition. Conclusions are drawn in Section 6.

2 Local Binary Patterns (LBP)

The original LBP operator was introduced by Ojala et al [11]. The operator labels the pixels of an image by thresholding a 3×3 neighbourhood of each pixel with the center value resulting in a binary number (see the left side of Fig 1). Then the histogram of the labels was used as a texture descriptor.

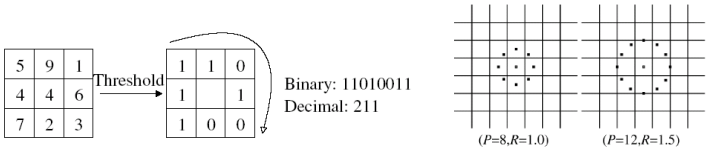

Figure 1: Left: The basic LBP operator [1]. Right: Two examples of the extended LBP [11]: a circular $(8,1)$ neighborhood, and a circular $(12,1.5)$ neighbourhood.

The small 3×3 neighbourhood of the basic LBP opterator can not capture dominant features with large scale structures. Hence the operator was extended to use neighbourhood of different sizes [11]. Using circular neighbourhoods and bilinearly interpolating the pixel values allows any radius and number of pixels in the neighbourhood. Examples of the extended LBP are shown in the right side of Fig 1, where (P,R) denotes P sampling points on a circle of radius of R. Further extension of LBP introduced uniform patterns [11]. A Local Binary Pattern is called uniform if it contains at most two bitwise transitions from 0 to 1 or vice versa when the binary string is considered circular.

Here we adopt the notation $LBP^{u2}_{P,R}$ for LBP operators: the subscript represents using the operator in a (P,R) neighbourhood, and the superscript $u2$ indicates using only uni-

form patterns and labelling all remaining patterns with a single label. A histogram of a labelled image $f_l(x,y)$ can be defined as

$$H_i = \sum_{x,y} I(f_l(x,y) = i), \qquad i = 0,\ldots,n-1 \tag{1}$$

where n is the number of of different labels produced by the LBP operator and

$$I(A) = \begin{cases} 1 & A \text{ is true} \\ 0 & A \text{ is false} \end{cases} \tag{2}$$

This histogram contains information about the distribution of the local micro-patterns, such as edges, spots and flat areas, over the whole image.

Face images can be seen as a composition of micro-patterns which can be effectively described by the LBP histograms. In [1], face images were equally divided into small regions to extract LBP features (see the left side of Fig 3 for an illustration). However, this LBP feature extraction scheme suffers from fixed LBP feature size and positions. Here we propose to learn discriminative LBP features using boosting learning from a large LBP features pool obtained by shifting and scaling a sub-window over face images.

3 AdaBoost

AdaBoost, introducted by Freund and Schapire [7, 12], provides a simple yet effective approach for stagewise learning of a nonlinear classification function. AdaBoost learns a small number of weak classifiers whose performance are just better than random guessing, and boosts them iteratively into a strong classsifier of higher accuracy. The process of AdaBoost maintains a distribution on the training samples. At each iteration, a weak classifier which minimizes the weighted error rate is selected, and the distribution is updated to increase the weights of the misclassified samples and reduce others' weights. AdaBoost has been successfully used in many problems such as face detection [16].

Here we apply AdaBoost to boost LBP-based weak classifiers. For weak classifier, we adopt template matching as follows. In training, the LBP histograms in a given class are averaged to generate a histogram template for this class. In recognition, a nearest-neighbour classifier is adopted: the input histogram is matched with the closest template. We select the Chi square statistic (χ^2) as the dissimilarity measure for histograms:

$$\chi^2(\mathbf{S},\mathbf{M}) = \sum_i \frac{(S_i - M_i)^2}{S_i + M_i} \tag{3}$$

where S and M are two LBP histograms.

However, Li and Zhang [10] have shown recently that a strong classifier learned by AdaBoost is suboptimal, and proposed FloatBoost by incorporating Floating Search into AdaBoost. FloatBoost uses a backtrack mechanism after each iteration of AdaBoost to remove those weak classifiers that are ineffective in reducing error rate. Compared to AdaBoost, FloatBoost is reported to yield a strong classifier consisting of fewer weak classifiers whilst improving classification performance. However, FloatBoost increases the training time massively compared to that for AdaBoost. In contrast, we proposed in the following a novel learning method to avoid selecting of ineffective weak classifiers in each iteration of learning but is also very fast in training.

4 Conditional Mutual Information based Boosting

Motivated by the Conditional Mutual Information (CMI) based binary feature selection proposed recently [6], we propose here a Conditional Mutual Information based Boosting (CMIB) scheme for effecient learning.

CMI based Feature Selection — Mutual Information (MI) is a basic concept in information theory. It estimates the quantity of information shared between random variables. For two random variabels U and V, their mutual information $I(U;V)$ is defined as follows:

$$I(U;V) = H(U) - H(U|V) = H(V) - H(V|U) \tag{4}$$

where $H()$ is the entropy of the random variable. The entropy $H(U)$ quantifies the uncertainty of U. For a discrete random variable U, $H(U)$ is defined as

$$H(U) = -\sum_{u \in U} p(u) \log p(u) \tag{5}$$

Here $p(u)$ represents the marginal probability distribution of U. The conditional entropy $H(U|V)$ quantifies the remaining uncertainty of U, when V is known.

Given M samples with the N features $X_1, \ldots, X_N$, and the target classification variable Y, feature selection is to find K features $X_{v(1)}, \ldots, X_{v(K)}$ that optimally characterizes Y. Mutual Information based feature selection is to select features $v(1), \ldots, v(K)$ which individually maximize the mutual information $I(Y;X_{v(l)})$.

However, selection based on such a criterion cannot ensure weak dependency among features, and can lead to redundant and poorly informative families of features. Recently Fleuret [6] proposed a Conditional Mutual Information (CMI) maximization criterion to select features. The essence is that a feature X can be discarded if there is one feature X_v already picked such that X and Y are conditionally independent given X_v. Conditional Mutual Information is defined as

$$I(U;V|W) = H(U|W) - H(U|W,V) \tag{6}$$

that measures the information shared between U and V when W is known. If V and W carry the same information about U, the two terms on the right are equal, and the CMI is zero, even if both V and W are individually informative. On the contrary if V brings information about U which is not already contained in W, the difference is large.

For feature selection, a feature X' is good only if $I(Y;X'|X)$ is large for every X already picked. This means that X' is good only if it carries information about Y, and if this information has not been caught by any of the X already picked. An iterative procedure for a CMI based feature selection can be defined as

$$v(1) = \arg\max_n I(Y;X_n) \tag{7}$$

$$\forall k, 1 \leq k < K, v(k+1) = \arg\max_n \left\{ \min_{l \leq k} I(Y;X_n|X_{v(l)}) \right\} \tag{8}$$

$I(Y;X_n|X_{v(l)})$ is small either if X_n contains no information about Y or if such information was already in $X_{v(l)}$. Note that the equivalent criterion was also proposed in [15].

CMI based Boosting (CMIB) — We propose to learn a small set of weak classifiers from a large classifier pool using CMI, and boost them into a strong classifier. We regard the output of a weak classifier as a random variable, a 'feature' for the candidate class; and employ the CMI maxizimation criterion to select the effective 'features', i.e. the characterizing weak classifiers. CMIB learns a sequence of weak classifiers which maximize their mutual information about a candidate class, conditional to the response of any weak classifier already selected. So a weak classifier similar to those that were already learned will not be selected, even if it is individually powerful as it does not carry additional information about the candidate class.

After learning weak classifiers, a strategy is needed to perform final classification by combining the learned weak classifiers. CMIB adopts the Naive-Bayes to make the final decision based on outputs of the weak classifiers, not the voting procedure used in AdaBoost. A Naive-Bayes classifier is simple but highly effective if the features can be assumed to be largely independent for a given class. As the weak classifiers learned by CMIB are by their very nature weakly dependent, it is reasonable to use the Naive-Bayes to combine them for final classification. If using c to represent the value of the class variable, and $x_1, \cdots, x_k$ for the features, a Naive Bayesian classifier is defined as

$$\hat{c} = \arg\max_{c} p(c) \prod_{i=1}^{k} p(x_i|c) \tag{9}$$

The proposed CMIB algorithm is summaried in Fig 2. CMIB learns weak classifiers that are both individually informative and weakly dependent. Crucially, CMIB avoids selecting unfavorable weak classifiers in each iteration, while FloatBoost [10] delects ineffective weak classifiers after each iteration.

AdaBoost usually requires very expensive training time. The improved performance of FloatBoost also pays the price for 5 times longer training time than AdaBoost [10]. In constrast, CMIB promises very fast training. The fast training is very significant for any incremental or adaptive learning when the size of the initial available training data is small but accumulating over time. (The fast training of CMIB will be illustrated in the follwing experiments.)

5 Experiments

Psychophysical studies indicate that basic emotions have corresponding universal facial expressions across all cultures. This is reflected by most current facial expression recognition systems [3, 13, 2] that attempt to recognize a set of prototypic emotional expressions including disgust, fear, joy, surprise, sadness and anger. In this section we evaluate performance on both 6-class prototypic expression recognition and 7-class expression recognition by including the neutral expression.

Dataset — The Cohn-Kanade Facial Expression Database [9] was used here. This database consists of 100 university students aged from 18 to 30 years, of which 65% were female, 15% were African-American, and 3% were Asian or Latino. Subjects were instructed to perform a series of 23 facial displays, six of which were prototypic emotions mentioned above. Image sequences from neutral to target display were digitized into 640×490 pixel arrays.

Given a training set $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ where $y_i \in \{-1, 1\}$ and the size of the final strong classifier T:

1. Train weak classifiers $H_1(x), \ldots, H_N(x)$ based on training samples, where N is the total number of weak classifiers
2. For $t = 1, \ldots, T$,
   - If $t = 1$, choose $H_t(x) = H_j(x)$, such that
   $$j = \arg\max_n I(Y; H_n(x)) \tag{10}$$
   - If $t > 1$, choose $H_t(x) = H_j(x)$, such that
   $$j = \arg\max_n \left\{ \min_{l<t} I(Y; H_n(x) | H_l(x)) \right\} \tag{11}$$
3. Output the final strong classifier as
$$f(x) = \arg\max_y p(y) \prod_{i=1}^{T} p(H_i(x)|y) \tag{12}$$

Figure 2: Conditional Mutual Information based Boosting.

For our experiments, we selected 320 image sequences from the database. The only selection criterion was that a sequence could be labeled as one of the six basic emotions. The sequences come from 96 subjects, with 1 to 6 emotions per subject. For each sequence, the neutral face and three peak frames were used. To evaluate generalization performance, a 10-fold Cross-Validation testing scheme was adopted.

Following Tian [13], we normalized the faces to a fixed distance of 55 pixels between the centers of the two eyes. It is observed that the width of a face is roughly twice this distance, and the height is roughly triple. Hence, facial images of 110×150 pixels were cropped from original frames based on the two eyes location. No further alignment of facial features such as alignment of mouth [17] was performed in our algorithms. Due to LBP's gray-scale invariance, there was no attempt made to remove illumination changes [13] in our algorithm.

Expression Recognition using LBP — Experiments were performed to evaluate the effectiveness of LBP features for facial expression recognition. Here a face image was equally divided into small sub-regions from which LBP features were extracted and concatenated into a single, spatially enhanced feature histogram [1]. The extracted histogram represents the local texture and global shape of face images. We divided 110×150 pixels facial images into 18×21 pixels regions giving 42 (6×7) sub-regions in total (as shown in Fig. 3). We adopted the 59-bin $LBP^{u2}_{8,2}$ operator for each sub-region. The length of the extracted histogram is 2478 (59×42).

We adopted template matching as the classifier for its simplicity. In training, the histograms of face images in a given class are averaged to generate a histogram template

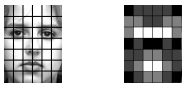

Figure 3: Left: A face image divided into 6×7 sub-region. Right: The weights set for weighted dissimilarity measure. Black squares indicate weight 0.0, dark gray 1.0, light gray 2.0 and white 4.0.

for this class. In recognition, a nearest-neighbour classifier is adopted: the histogram of the input image is matched with the closest template.

For dissimilarity measure, we selected Chi square statistic (χ^2) described in Section 3. It is observed that facial features contributing to facial expressions mainly lie in regions such as eye and mouth regions. These regions contain more useful information for expression classification. Therefore, a weight can be set for each region based on its importance, as shown in Fig 3. This particular weight set was designed empirically based on observation. Our weighted (χ^2) statistic is then given as

$$\chi_w^2(\mathbf{S},\mathbf{M}) = \sum_{i,j} w_j \frac{(S_{i,j} - M_{i,j})^2}{S_{i,j} + M_{i,j}} \tag{13}$$

where S and M are two LBP histograms, w_j is the weight for region j.

The template matching achieved the generalization performance of 84.5% and 79.1% for the 6-class task and the 7-class task respectively. Based on the tracked geometric facial features (eyebrows, eyelids, and mouth), Cohen et al [3] adopted Bayesian network classifiers to classify 7-class emotional expressions on the Cohn-Kanade database. The best performance of 73.2% was obtained by them using Tree-Augmented-Naive Bayes (TAN) classifiers. Comparison in Table 1 illustrates that our simple template matching using LBP outperforms geometric features based TAN classifier. The experiments demonstrated that the low-cost LBP features are discriminative for facial expression recogniton.

Methods (Feature + Classifier)	Recogniton Results
LBP + Template Matching	79.1%
Geometric Feature + TAN [3]	73.2%

Table 1: Comparisons between the geometric features based TAN [3] and our LBP-based template matching.

Expression Recognition using Boosted LBP —By shifting and scaling a sub-window, 16,640 LBP features in total were extracted from each face image. We adopted CMIB and AdaBoost to learn a small subset (in tens) of effective LBP features, and then recognize facial expressions using the boosted strong classifiers. CMIB and AdaBoost used the same weak classifier described in Section 3. For AdaBoost, we used the generailized multi-class multi-label AdaBoost.MH algorithm proposed in [12].

Training Computational Complexity: We plot the average training time of CMIB and AdaBoost as a function of the number of weak classifiers in Fig 4. The experiments were

run on a standard 2.0Ghz PC with Matlab implementation. It can be seen that CMIB performs significantly fast than AdaBoost, especially when the number of learned weak classifiers increases. For example, CMIB selects top 100 weak classifiers with an average time of 1.5t, while AdaBoost needs 38.3t for the same task. (The varation in AdaBoost running time was due to network load and system load, since we conducted experiments with Matlab installed in a central server. Even with such variation, we can still observe that the training time of AdaBoost is linear in number of rounds.)

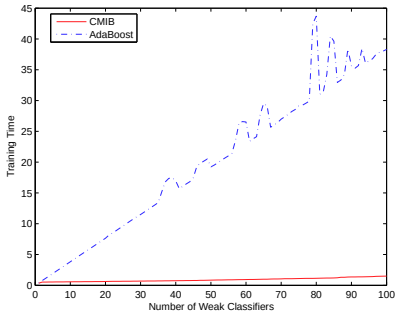

Figure 4: Training time of CMIB and AdaBoost, as a function of the number of weak classifiers. Up to 100 weak classifiers are considered.

Classification Accuracy: We conducted expression recognition using the strong classifiers boosted by CMIB and AdaBoost. The generalization performance in 6-class and 7-class recognition are shown in Fig 5, as a function of the number of weak classifiers. Here a strong classifier is composed of up to 200 weak classifiers.

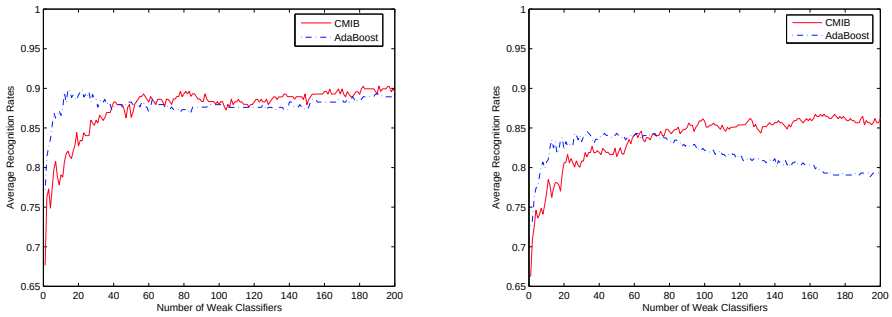

Figure 5: Generalization performance of 10-fold Cross-Validation, as a function of the number of weak classifiers. Left: **6-class**; Right: **7-class**.

The following conclusions can be drawn from the results shown in Fig 5: (1) The generalization performance is clearly improved for both the 6-class and 7-class recognition tasks by boosting LBP-based classifiers over that of LBP without boosting (i.e., with uniformly divided sub-regions). (2) CMIB achieves as good or better recognition results than

Methods	Recogniton Results
Boosting LBP with CMIB	86.7%
Boosting LBP with AdaBoost	84.6%
Boosting Gabor wavelets with AdaBoost [2]	85.0%

Table 2: Comparisons between the Gabor-wavelets-based boosting [2] and our LBP-based boosting.

AdaBoost, though AdaBoost performs better when using less than 40-60 weak classifiers. CMIB performs consistently better when more weak classifiers are learned. Bartlett et al [2] performed similar experiment on the Cohn-Kanade database using AdaBoost to learn Gabor-wavelet features. Comparisons summaried in Table 2 illustrate that boosting LBP with CMIB performs better, while boosting LBP with AdaBoost performs comparably to boosting Gabor-wavelet features. (3) CMIB improves its recognition performance when the number of weak classifiers increases, while AdaBoost may decrease its recognition rates with more weak classifiers learned. Since CMIB avoids ineffective weak classifiers in learning, more weak classifiers will produce better recognition performance. In contrast, the performance of the classifier boosted by AdaBoost may degenerate when adding the unfavorable weak classifiers.

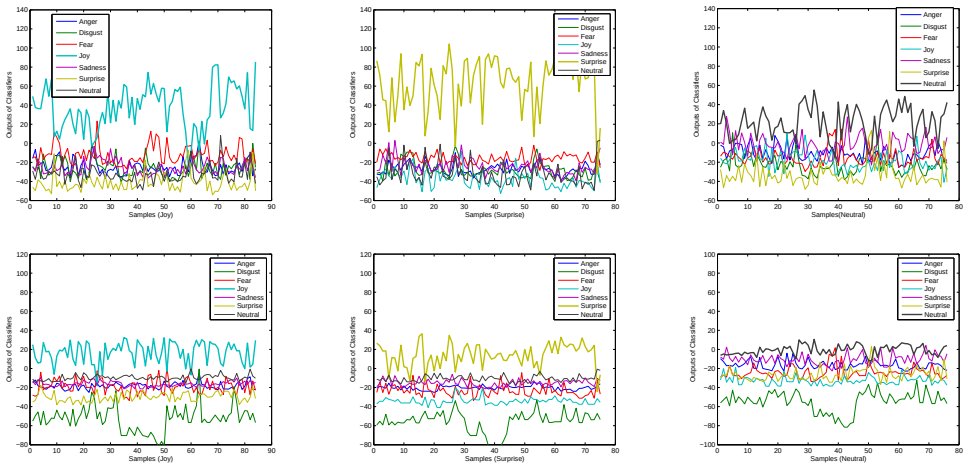

Figure 6: Outputs of classifiers for samples Joy, Surprise, and Neutral. The first row: **CMIB**; the second row: **AdaBoost**.

Between-Class Discriminative Robustness: Due to the limitation of space, we only show here the outputs of boosted classifiers of different expressions for samples "Joy", "Surprise", and "Neutral" in Fig 6. It can be seen that the different weak classifiers being learned by CMIB and AdaBoost have some impact on not only the recognition accuracy, but also the robustness of recognition. Weak classifiers learned by CMIB provide better discriminative ability in between-class seperation than that of AdaBoost, resulting in more robust recognition.

6 Conclusions

This paper presented a novel method for facial expression recognition by boosting Local Binary Patterns (LBP) based classifiers. Low-cost LBP features were introduced to effectively describle appearance features of expression images. A novel learning procedure, Conditional Mutual Information based Boost (CMIB), was proposed for efficient learning. Extensive experiments illustrated that LBP features are effective for expression analysis, and CMIB is superior to AdaBoost in training complexity and classification performance.

References

[1] T. Ahonen, A. Hadid, and M. Pietikinen. Face recognition with local binary patterns. In *ECCV*, 2004.

[2] M.S. Bartlett, G. Littlewort, I. Fasel, and R. Movellan. Real time face detection and facial expression recognition: Development and application to human computer interaction. In *CVPR Workshop on CVPR for HCI*, 2003.

[3] I. Cohen, N. Sebe, A. Garg, L. Chen, and T. S. Huang. Facial expression recognition from video sequences: Temporal and static modeling. *CVIU*, 91:160–187, 2003.

[4] G. Donato, M. Bartlett, J. Hager, P. Ekman, and T. Sejnowski. Classifying facial actions. *IEEE PAMI*, 1999.

[5] B. Fasel and J. Luettin. Automatic facial expression analysis: a survey. *Pattern Recognition*, 36:259–275, 2003.

[6] F. Fleuret. Fast binary feature selection with conditional mutual information. *Journal of Machine Learning Research*, pages 1531–1555, 2004.

[7] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[8] A. Hadid, M. Pietikinen, and T. Ahonen. A discriminative feature space for detecting and recognizing faces. In *CVPR*, 2004.

[9] T. Kanade, J.F. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. In *IEEE FG*, 2000.

[10] S. Z. Li and Z. Zhang. Floatboost learning and statistical face detection. *IEEE PAMI*, 2004.

[11] T. Ojala, M. Pietikinen, and T. Menp. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE PAMI*, 2002.

[12] R. E. Schapire and Y. Singer. Improved boosting algorithms using confidence-rated predictions. *Maching Learning*, 37(3):297–336, 1999.

[13] Y. Tian. Evaluation of face resolution for expression analysis. In *CVPR Workshop on Face Processing in Video*, 2004.

[14] Y. Tian, T. Kanade, and J.F. Cohn. Facial Expression Analysis, in *Handbook of Face Recognition*, Springer, 2003.

[15] M. Vidal-Naquet and S. Ullman. Object recognition with informative features and linear classification. In ICCV, 2003.

[16] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, 2001.

[17] Z. Zhang, M. J. Lyons, M. Schuster, and S. Akamatsu. Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron. In *IEEE FG*, 1998.